International Journal of Research, Science, Technology & Management
ISSN Online: 2455-2240, Volume 15 Issue I, April 2019

GEOSTATISTICAL ANALYSIS RESEARCH USING ADULT DATA SET

Ritu Kumar, Shraddha Desai

ABSTRACT

With the great progress of microelectronics and other relating information technologies, together with the still broadening applications of computers in a vast range of businesses and industries, large databases containing mixed-mode data are becoming quite commonplace. Today, large databases contain various modes of collected data related to different components of a complex real world system. Their use is not necessarily confined to classifications. Many of them may not have clearly-defined class labels or even any explicit class information at all. Indeed, there are many different reasons to determine or discover all patterns, to achieve any comprehensive analysis and understanding of the information within the data spaces. In the past, data mining or pattern discovery has by and large been developed fundamentally for categorical databases. All of the classification rules have been found from pre-labeled data samples. For a large mixed-mode database, how to discredit its continuous data into interval events is still a practical approach. If there are no class labels for the database, we have no helpful correlation references to such task actually a large relational database may contain various correlated attribute clusters. To handle these kinds of problems, we first have to partition the databases into sub-groups of attributes containing some sort of correlated relationship. This process has become known as attribute clustering, and it is an important way to reduce our search in looking for or discovering patterns Furthermore, once correlated attribute groups are obtained, from each of them, we could find the most representative attribute with the strongest interdependence with all other attributes in that cluster, and use it as a candidate like a class label of that group. That will set up a correlation attribute to drive the discretization of the other continuous data in each attribute cluster. This thesis provides the theoretical framework, the methodology and the computational system to achieve that goal.

In validating the premises proposed in the Paper, extensive experiments using UCI Expository Data of various types were performed to verify each of the fine points conceived. To demonstrate the usefulness for solving real world problems, the developed methodology is applied to Adult databases from the real world.

Keywords: Pattern Discovery, Mixed mode Database, discretization

1. INTRODUCTION

In the past decade, with the development of semiconductors, microelectronics, cloud processors, magnetic storage media and other information acquisition methods, together with the continually broadening applications of computers in a wide range of businesses and industries, large databases have become quite a commonplace. The volumes of these databases have been growing from megabytes to gigabytes, and to terabytes and even to petabyte. The types of database they contain also vary: some of them could be numeric, others could be categorical and the most common ones are a mixture of both. These are referred to as “mixed-mode databases”.

Today we are facing large relational databases with mixed-mode attributes. Many of those have either no class labels, or no defined class information. They may contain different modes of correlated data, related to different attributes of a complex system. Their uses are not confined to classification. Nevertheless, there is a great need for discovering patterns among them for comprehensive analysis, interpretation and understanding the patterns or relations inherent in the data. Analyzing these kinds of mixed-mode databases, and thus supplying decision-makers with useful knowledge, is very challenging. Developing new measurements to transfer data into knowledge bases is now a paramount problem in data mining research community. The objective of this thesis is to develop methods for discovering patterns in mixed-mode data where class information is non-existing or unavailable.

Two important issues that have to be discussed for partitioning continuous data are the number of intervals and the ranges of the relative intervals. These two problems must be solved, either by the discretization algorithm itself or by designation by the engineers [17]. Most partition algorithms require the provision of the suitable number of data intervals by the engineers. The widths of the intervals can be defined too, through the boundaries of the discretized data intervals. A good algorithm for this purpose should usually require only a few input parameters from the operators. As a specific

real-world classification problem, the available class information could provide crucial support to the discretization process. The rest of the remaining problem consists of how to partition a continuous database with continuous attributes in a mixed-mode data set [12].

1.2 The Motivation

The challenges are enumerated below:

1) With respect to the structures of mixed-mode data space, they are becoming more complicated than ever before. The data values could be mixed-mode, consisting of both categorical data and continuous data. At the same time, the data dimensionality could be huge. Data could have been gathered systematically over a long period, and be piled up more-or-less randomly.

2) With respect to the quality of the mixed-mode data space, undoubtedly there are many reasons and causes in the real world for data to be collected affected by many kinds of noise. Here, probabilistic approaches must be implemented in real-world databases, instead of deterministic approaches

3) With respect to applying useful patterns discovered in real production processes, a certain kind of measurement for pattern confidence and support should be implemented to render reliable data pattern analysis results and assist in the decision-making process.

4) With respect to a priori real domain knowledge, in most situations it is difficult or even impossible to collect adequate domain knowledge for effective decision-making. This is definitely the truth for investigations in some new application fields. Some of special domain experts, who can support some observations and measurements to set up a domain database, but will expect to get some suggestions or evidences for data analysis results for realizing and even formulating theoretical or operational ideas.

Problems encountered in mixed-mode database

1. Possible existence of unknown attribute-interdependent groups

For a large database with a large number of mixed-mode attributes, it is possible that several strongly attribute-correlated groups may exist within one data space. They could finally be found if we have an attribute clustering algorithm to do the job. In classical pattern recognition and data mining procedures, clustering is an important issue. Given a relational table, any of the conventional clustering algorithms will cluster tuple into several groups, each of which is characterized by a set of attribute values based on similarity [16]. Intuitively, tuples in a cluster are more similar to each other within the same cluster than those belonging to different clusters.

It has been shown that clustering is very helpful in many data mining tasks. In the past clustering methods are mostly developed to group samples. However, a majority of the pending problems is the data set has too many attributes which might not even be correlated. To perform pattern discovery on a large mixed-mode database, this dissertation presents a new methodology to group mixed-mode attributes that are interdependent and/or correlated to each other instead. We refer to such a process as “attribute clustering”. In this sense, attributes in the same cluster are more correlated to each other, whereas attributes in different clusters are less correlated.

2. Attribute clustering before discretization of continuous data

Following the observation in the last paragraph, this dissertation will present an attribute clustering method which is able to group mixed-mode attributes within the database automatically, based on their interdependence, so that meaningful patterns can be discovered later. The partitioning of a relation database into attribute subgroups produces a small number of attributes, within and then across the groups, to be defined for data mining tasks. After attribute clustering, the search dimensionality of each dataset for a data mining algorithm is reduced significantly [35]. The reduction of search dimensionality is especially important for data mining in very large mixed-mode databases, particularly in databases consisting of a huge number of attributes and a small number of samples. The situation could become even worse when the number of attributes overwhelms the number of tuples.

In such cases, the patterns discovered that are actually random becomes rather higher than the usual situation. It is for the abovementioned reasons that attribute grouping is an important pre-processing stage for many data mining algorithms, to ensure effectiveness when applied to a very large database [32].

2. REVIEW OF LITERATURE

2.1 Pattern Discovery

Pattern discovery, as one of the powerful intelligent decision support platforms, is being increasingly applied to large-scale complicated systems and domains even in mixed-mode space [23]. It has been shown that it has the capacity to extract useful knowledge from a large data space and present to the decision makers. It is growing gradually and becomes more important with the quick development of computer technologies with increasing capacity to collect massive

amounts of valuable data for pattern analysis. Extracting relevant information and useful knowledge from large mixed-mode data spaces is still complicated by several challenging issues: the limitations of data storage formats; a lack of expert prior knowledge for real-world databases; the difficulty of visualizing the data using inefficient data mining tools. Data mining is a series of steps in the knowledge discovery process, consisting of the use of particular algorithms for producing patterns, as required by the real world. Useful information being extracted from real-world data using traditional data mining tools may be made better by the prior perception of a domain knowledge base or expert experience. One could use the classical data mining tools [32] to get supporting data to confirm or refute existing personal perception, but one also cannot be assured that there are no better-fitting explanations for the discovered patterns, or even that no important information has been missed in the entire data mining process. For a relatively complex real problem with a large data space, all traditional knowledge acquisition and data mining tools would become obviously inefficient, even helpless in some ways.

In decision-support, it is very easy to be biased by the subjectivities of the domain experts, or even by pre-assumptions used in data mining and the algorithmic procedures thereof. While most of the current approaches are trying to combine decision trees, neural network technologies, and the like, for pattern discovery and decision support, the rationale is to have a systematic solution providing decision-making procedure or predictive rules derived from the patterns inherent in the data space. Regarding most of the existing data mining systems, some of the accessory processes like pre-processing, data cleansing, filtering, attribute reduction are proposed [27] in order to remove data noise by bringing out more relevant information from the data space, and to reduce the search space and time, and thus cost, for that procedure.

All of the approaches discussed above make researchers investigate patterns and then verify the classification by domain experts, who often depend on their prior knowledge - including the parameters of the predetermined systematic classification framework. In that situation, they may be biased, and usually have to make long iterative search activities with personal examination and re-examination routine procedures. Due to the limited personal abilities to explore new patterns and knowledge, it is often difficult to set up a more objective base for decision-making. For a larger mixed-mode database with more unanticipated variations than normal ones, even the domain experts would find it difficult to reach useful results [27]. Furthermore, in the real world, three other important topics must be faced by the decision-makers, these being: 1) flexibility and versatility of the pattern discovery procedure; 2) transparency to get at supporting evidence; and finally 3) the processing cost and computation speed.

In conclusion, if the tools for pattern discovery could be easily implemented by the real world users, those tools should have the following basic characteristics: 1. Discover multiple patterns from a data space without relying on prior knowledge as supporting evidence;

2. Collaborate with flexible decision objectives and situations.

3. Provide significant discovered patterns for the following analysis.

4. Render a reconstruction framework with high speed of computation at low cost.

To satisfy these important and basic needs, a new pattern discovery approach has to be developed [22], which should be a primarily data-driven one. To discover an unbiased and statistically significant event automatically and exhaustively is now feasible. From the discovered patterns, classification modules for categorization and prediction can now be realized. At least one unique feature of the potential system is the ability to discover multiple significant patterns of high order at very fast speed, and then to list them according to their statistical confidence levels, so that a better understanding of the pattern and rules can be achieved [22]. Based on this theoretical and systematic framework design, a software platform has been developed along with several new feature modules including attribute clustering [17], class-dependent discretization [15]; classification and forecasting [43]. In this dissertation, the main emphasis has been on overcoming the difficulties in handling mixed-mode data in the new theoretical framework, and on demonstrating the performance of the new platform, especially when applying to large databases from real-world problems.

Those very initial research activities began in the early seventies by Wong [15] who first attempted to explore for quantitative information measurements and statistical patterns in English text [22], and then in digital image databases [24]. With the strong belief that information in bio-molecular data sequences is coded for bio-molecular structures, he has made a great effort to calculate quantitative information measurements and statistical patterns discovered in the bio-molecules database. It has been proved that statistical patterns discovered which present the underlying biochemical and taxonomical features, can be identified and then analyzed later. Following up this line of thought, information on quantitative measurements of how the data deviated from equal-probability and also independence models has been set up for English texts analysis [27] and

images understanding [26]. These important discoveries finally formed the early basis of today's pattern discovery approach, as discussed in this thesis. Pattern recognition algorithms for discrete continuous data space were well developed later for other real applications [28].

More recent research has noted that if the dimensionality of a real mixed-mode database is very large, this will make the definition of patterns discovered within the traditional pattern discovery framework much less meaningful [29]. Although various pattern discovery methods have been developed [44], they all depend on the interdependency of attributes with the consideration of attributes as the random variables. In fact, all of the higher order pattern discovery platforms have been developed [45] only for discrete databases. Within those discovery frameworks, patterns have been defined as statistically significant associations of two or more primary events from different attributes in the analysis data space. For exploring patterns in databases in the presence of data noise, we have developed the adjusted residual analysis approach, which guarantees that the discovered patterns are not resulting from random association, with a fixed confidence level. All of the high-order patterns discovered can then be applied to support application tasks such as classification or pattern clustering. At the same time, the entire high-order pattern discovered within the continuous database was also advanced. Events here for the continuous data space are defined as Borel sets [45] and thus the pattern discovery is transferred into an optimization problem of finding the hypercells such that the frequency of data p oints if contains deviate statistically significantly from the default space-wise uniformity model. Analysis tasks, including classification and probability density estimation, will be easily performed based on the patterns discovered, as well as the significant analysis results on both artificial and real-world databases have been completed.

These automatic pattern discovery algorithms become a good and helpful platform to support different types of decision-making tasks in the real world. As reported in [45]. while good solution to discover patterns and construct non-parametric probability density in continuous data space with scale invariant properties has been developed, the scalability for this approach is questionable because the hyper cells for defining g high order significant statistical events is built on the genetic algorithm.

3. EXPERIMENT RESULTS

Adult Data Set (Mixed Mode Data)

This database was extracted from the census bureau database found at (Table 1).
It contains 48842 instances of mix of continuous and discrete data with 14 attributes (Table 2). It has been used for predictive whether a person makes over 50k a year or not. We use this mixed-mode data set to answer the questions (a) to (e). More specifically, the experiment is used: 1) to demonstrate the existence of attribute subgroups in the mixed-mode data set; 2) to illustrate the attainment of attribute cluster configuration and the grouping of cluster items in situations with or without class label; 3) to show the classification characteristics of various attributes in different attribute groups found by ACA; 4) to show that the attribute with highest normalized SR, or simply the mode, in the attribute group is usually with high classification rate if it is assumed to take the role of a class label. The experiment results show that the mode in each attribute group/cluster can be considered as the most discriminative/representative or governing attribute to drive the discretization of continuous attributes in the attribute group/cluster.

In this experiment, the proposed method is used to calculate the normalized mutual information, *R*, among the attributes. Their values are tabulated in Table 3 for the dataset with class label excluded and in Table 4 with class label included. Based on the *R* values, our ACA found the optimal cluster configuration in the given data set.

Table 1 A Brief Description of Adult Data Set
Data Description

Data Set	Attribute Characteristics	No. of Samples	No. of Attributes	No. of Classes
Adult	Mixed-Mode Data	48842	14	2

Table 2 the Attributes of Adult Data Set

Attribute	Name	Characteristics
*A*1	Work class Discrete	Discrete
*A*2	Education	
*A*3	marital-status	
*A*4	Occupation	
*A*5	Relationship	
*A*6	Race	
*A*7	Sex	

A8	native-country	
A9	Age	Continuous
A10	Fnlwgt	
A11	Education-num	
A12	capital-gain	
A13	capital-loss	
A14	Hours per week	
Class	income	Discrete

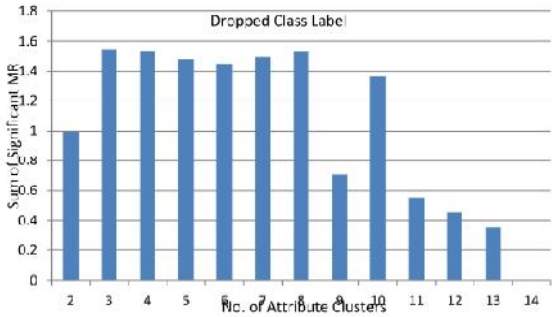
In our proposed method, no class information is required; nevertheless, the results reported in Table 6 shows that even without class information, our proposed method and ACA are able to group interdependent attributes together. This demonstrates the effectiveness of our method to extract the same intrinsic information inherent in the classes.

Table 3 the Sum of Significant MR obtained for each k of the k-Mode ACA

No. of Attribute Cluster, k	Excluded Class Label Sum of Significant MR	Included Class Label Sum of Significant MR
2	0.993065	1.559977
3	1.546628	1.599815
4	1.536047	1.597278
5	1.478268	*1.685009
6	1.452389	1.603821
7	1.498127	1.504654
8	1.53544	1.522862
9	0.708419	1.032613
10	1.366747	1.393317
11	0.553327	0.914691
12	0.452937	0.553327
13	0.355844	0.452937
14	0	0.763257
15	-	0

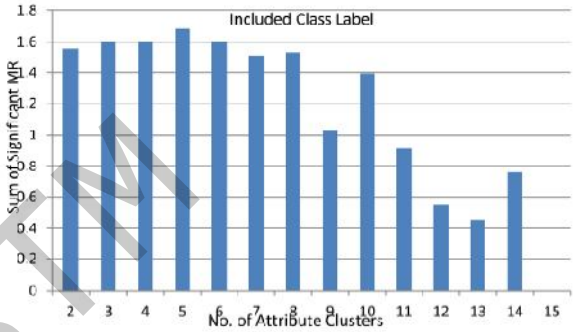
* Highest Sum of Significant MR Implies Optimal k =3 for Data Set Dropped Class Label and Optimal k = 5 for Data Set Included Class Label.

Table 4 the Plot of the Sum of Significant MR (with class label dropped)



* Highest Sum of Significant MR Implies Optimal k =3.

Table 5 the Plot of the Sum of Significant MR (with class label included)



* Highest Sum of Significant MR Implies Optimal k =5.

Table 6 The Attribute Clusters and their Mode Obtained by ACA

Attribute Cluster Items		
Attribute Group	**Dropped Class Label**	**Included Class Label**
1	*native-country, race, fnlwgt	*native-country, race, fnlwgt
2	*education, workclass, occupation, education-num	*education-num, education
3	*relationship, marital-status, sex, age, capital-gain, capital-loss, hours-per-week	*relationship, marital-status, sex, age
4	-	*workclass, occupation
5	-	*income (class), capital-gain, capital-loss, hours-per-week

* The attribute marked with "*" is the mode of the attribute group. A mode is with the highest normalized mutual information in the attribute group.

To further investigate the attributes resided in each attribute group, we study the classificatory aspect of them to show that in a normal setting the mode is also the attribute that renders good enough classification rate if it is regarded as a class label. The attribute clusters, normalized SR values and their classification performance are tabulated in Table 7.

Table 7 Attribute Clusters of Adult Data with Class Label Excluded with their Normalized SR Values and their Classification Accuracy by PD with a 95% Confidence

Attribute	Characteristics	Normalized SR	Classification Accuracy (%)
* native-country	Discrete	0.0952	89.59
race	Continuous	0.0898	84.43
fnlwgt	Continuous	0.0083	5.41

* The attribute marked with "*" is the mode of the attribute group. A mode is with the highest normalized mutual information in the attribute group.

Attribute	Characteristics	Normalized SR	Classification Accuracy (%)
*education	Discrete	0.8263	71.09
workclass	Discrete	0.8218	57.69
occupation	Discrete	0.2051	20.94
Education-num	Continuous	0.1173	-

* The attribute marked with "*" is the mode of the attribute group. A mode is with the highest normalized mutual information in the attribute group.

Attribute	Characteristics	Normalized SR	Classification Accuracy (%)
*Relationship	Discrete	0.6251	72
Martial status	Discrete	0.5525	74.78
Sex	Discrete	0.2465	68.95
age	Continuous	0.2229	-
^ capital-gain	Continuous	0.1100	99.51
^ capital-loss	Continuous	0.0495	95.33
hours-per-week	Continuous	0.0313	14.54

* The attribute marked with "*" is the mode of the attribute group. A mode is with the highest normalized mutual information in the attribute group. ^ The attribute marked with "^" implies the data
is sparse. # the attribute marked with "#" holds the highest classification accuracy, even higher than the mode.

4. CONCLUSION AND FUTURE RESEARCH

The research presented in this dissertation was motivated by the challenges we are confronting today: (1) an increasingly huge amount of raw mixed-mode data today require effective pattern discovery methods to unveil inherent subtle information for better understanding; (2) the pressing need to develop intelligent systems which are able to support knowledge discovery and decision support from overwhelming volume of discovered patterns; (3) the increasing demand of applications of discovered patterns in scientific, business and industry; and (4) the application limitation of most existing systems which are not general enough to solve problems on mixed-mode databases with numerous real-world applications.

The research works presented in this thesis have provided an integrated, flexible and generic framework for pattern discovery and analysis of large mixed-mode databases. Its applications cover databases with continuous, categorical and mixed- mode data. The validity and the effectiveness of the proposed methods have been backed by a number of successful experimental results. Their usefulness in real world applications has been demonstrated by the intriguing and revealing results obtained when applying to two large mixed-mode databases. One consists of a large set of meteorological data taken from a geographic area in Southern China and another is a set of massive multi-senor data taken from a delay coking plant.

REFERENCES

[1] A. K. C. Wong and G. C. L. Li, "Association Pattern Analysis for Pattern Pruning, Pattern Clustering and Summarization", to appear in Journal of Knowledge and Information Systems, 2010.

[2] AKC Wong and G Li, "Simultaneous Pattern Clustering and Data Grouping", IEEE Trans Knowledge and Data Engineering, Vol. 20, No. 7, pp 911-923, 2008.

[3] G Li and AKC Wong, "Pattern Distance Measures in Categorical Data for Pattern Pruning and Clustering", submitted to IEEE Trans. on Knowledge and Data Engineering.

[4] L Liu, AKC Wong, and Y Wang, "A Global Optimal Algorithm for Class-Dependent Discretization of Continuous Data", Intelligent Data Analysis, Vol 8, no 2, pp 151-170, 2004

[5] AKC Wong and Y Wang, 'Pattern Discovery: A Data Driven Approach to Decision Support,' IEEE SMC, Vol 33, no. 3, pp. 114-124, 2003.

[6] Y Wang and AKC Wong, "From Association to Classification: Inference Using Weight of Evidence,' IEEE Trans On Knowledge Systems, Vol 15, no 3, pp 914-925, 2003

[7] T Chau and AKC Wong, 'Pattern Discovery by Residual Analysis and Recursive Partitioning,' IEEE Trans on Knowledge and Data Engineering, pp 833-854, 1999

[8] AKC Wong, and Y Wang, 'High-Order Pattern Discovery from Discrete-Valued Data,' IEEE Trans On Knowledge Systems, pp 877-893,Vol 9, No 6, 1997

[9] JY Ching, AKC Wong, and Chan, KCC, "Class-dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data", IEEE PAMI, Vol 17, No 7, pp 641-651, July 1995

[10] K K Durston, D KY Chiu, A KC Wong and G CL Li, "Inferring higher-order structure in protein sequences: a granular computing analysis" submitted to BMC Genomics.

[11] AKC Wong, WH Au and KCC Chan, "Discovering High-Order Patterns of Gene Expression Levels", Journal of Computational Biology, Vol. 15, No.6, 2008. revision, 2008

[12] WH Au, KCC Chan, AKC Wong and Y Wang, "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data" IEEE/ACM Trans on Computational Biology and Bioinformatics, Vol 2, No2, pp 83-101, 2005.

[13] A.K.C. Wong, Information Pattern Analysis, Synthesis and Discovery, Chapter 7, pages 254–257. University of Waterloo, 1998.

[14] A. K. C. Wong and Y. Wang, "High Order Pattern Discovery from Discrete-Valued Data," IEEE Trans. on Knowledge and Data Eng., vol. 9, no. 6, pp. 877-893, 1997.

[15] A.K.C. Wong and Y. Wang, "Pattern Discovery: A Data Driven Approach to Decision Support," IEEE Trans. on Syst., Man, Cybern. – Part C, vol. 33, no. 1, pp. 114-124, 2003.

[16] A.K.C. Wong, D.K.Y. Chiu and W. Huang, 'A Discrete-Valued Clustering Algorithm with Applications to Bimolecular Data,' Information Sciences, vol. 139, pp. 97-112, 2002.

[17] A. K. C. Wong and D. K. Y. Chiu, "Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 9, no. 8, pp. 796-805, 1987

[18] A.K.C. Wong and C.C. Wang. DECA - a discrete-valued data clustering algorithm. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1(4):342–349, 1979. 27

[19] A.K.C. Wong and T.S. Liu, 'Typicality, Diversity and Feature Patterns of an Ensemble,' IEEE Trans. on Computers, vol. 24, no. 2, pp. 158-181, 1975

[20] A.K.C. Wong, T.S. Liu, and C.C. Wang, "Statistical Analysis of Residue Variability in Cytochrome C," Journal of Molecular Biology, vol. 102, pp. 287-295, 1976.

[21] Wai-Ho Au, Keith C.C. Chan, Andrew K.C. Wong, and Yang Wang, Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 2, NO. 2, APRIL-JUNE 2005".

[22] C. C. Wang and A. K. C. Wong, "Classification of Discrete-Valued Data with Feature Space Transformation," IEEE Trans. on Automatic Control, vol. AC-24, no. 3, pp. 434–437, 1979.

[23] D. Chiu, A. Wong, and B. Cheung. Information discovery through hierarchical maximum entropy. Journal of Experimental and Theoretical Artificial Intelligence, 2:117–129, 1990.

[24] J. Catlett. On changing continuous attributes into ordered discrete attributes. In Y. Kodratooe, editor, Proc. 5th European Working Session on Learning, pages 164–178, Porto, Portugal, 1991, March. Springer-Verlag Heidelberg.

[25] J. Y. Ching, A. K. C. Wong, and K. C. C. Chan, "Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data," IEEE Trans. on Pattern

Analysis and Machine Intelligence, vol. 17, no. 7, pp. 631-641, 1995.

[26] P. Clark and R. Boswell. Rule induction with CN2: Some recent improvements. In Proc. Fifth European Working Session on Learning, pages 151–163, Berlin, 1991. Springer.

[27] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In International Conference on Machine Learning, pages 194–202, San Francisco, CA, 1995.

[28] Usama M. Fayyad and Keki B. Irani. On the handling of continuous-valued attributes in decision tree generation. Machine Learning, 8(1):87–102, 1992.

[29] K. M. Ho and P. D. Scott. Zeta: A global method for discretization of continuous variables. In D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors, Knowledge Discovery and Data Mining, pages 191–194, Menlo Park, 1997. AAAI Press.

[30] Robert C. Holte. Very simple classification rules perform well on most commonly used datasets. Machine Learning, 11(1):63–90, 1993.

[31] Quinlan J.R. Induction of Decision Trees. Machine Learining 1, 1986.

[32] Quinlan J.R. C4.5:programs for Machine Learning. Morgan Kaufmann, 1993.[12] Tou J.T and Gonzalez R.C. Pattern Recognition Principles. Addison-Wesley, 1974.

[33] J.Y.Ching, A.K.C.Wong, and K.C.C.Chan. Class-dependent discretization for inductive learning from continuous data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(7):641–651, 1995.

[34] R. Kerber. Chi merge: Discretization of numeric attributes. In Proceedings of the 9th International Conference on Artificial Intelligence, pages 123–128, Menlo Park CA, 1992.

[35] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger. Mlc++: A machine learning library in c, 1994.

[36] T Kohonen. Self-Organization and Associative Memory. Springer-Verlag, Berlin, Germany, 1989.

[37] L. Kurgan and K.J. Cios. Discretization algorithm that uses class-attribute interdependence maximization. In Proceedings of the 2001 International Conference on Artificial Intelligence (IC-AI 2001), pages 980–987, Las Vegas, Nevada, 2001,JUNE.

[38] P. Langley. Induction of recursive bayesian classifiers. In P. Brazdil, editor, ECML93, volume 667 of LNAI, pages 153–164, Berlin, 1993. SV.

[39] Pat Langley, Wayne Iba, and Kevin Thompson. An analysis of bayesian classifiers. In National Conference on Artificial Intelligence, pages 223–228, San Jose, California, 1992.

[40] Huan Liu and Rudy Setiono. Feature selection via discretization. IEEE Transactions on Knowledge and Data Engineering, 9(4):642–645, 1997.

[41] P.M. Murphy and D.W. Aha. Uci repository of machine learning databases, 1994.

[42] A. Paterson and T.B. Niblett. Acls manual. Technical report, Intelligent Terminals Ltd., Edinburg, 1987.

[43] Bernhard Pfahringer. Compression-based discretization of continuous attributes. In International Conference on Machine Learning, pages 456–463, San Francisco, CA, 1995.

[44] J. R. Quinlan. Simplifying decision trees. International Journal of Man-Machine Studies, 27(3):221–234, 1987.

[45] J.R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kauffman, San, Mateo CA, 1993.

[46] M. RIcheldi and M. Rossotto. Class-driven statistical discretization of continuous attributes (extended abstract). Springer-Verlag, Berlin, Heidelberg, 1995.

[47] Moshe Sniedovich. Dynamic Programming, chapter appendix, pages 348–350. New York, 1992.

[48] R. Agrawal, S. Ghost, T. Imielinski, B. Iyer, and A. Swami, “An Interval Classifier for Database Mining Applications,” in Proc. of the 18th Int’l Conf. on Very Large Data Bases, Vancouver, British Columbia, Canada, 1992, pp. 560–573.

[49] R. Agrawal, T. Imielinski, and A. Swami, “Mining Association Rules between Sets of Items in Large Databases,” in Proc. of the ACM SIGMOD Int’l Conf. on Management of Data, Washington D.C., 1993, pp. 207–216.

[50] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” in Proc. of the 20th Int’l Conf. on Very Large Data Bases, Santiago, Chile, 1994, pp. 487–499.